Fast Sparse Raman Spectral Unmixing for Chemical Fingerprinting and Quantification

Fast Sparse Raman Spectral Unmixing for Chemical Fingerprinting and Quantification

Mehrdad Yaghoobi[a], Di Wu[a], Rhea J. Clewes[b], and Mike E. Davies[a]

[a]Institute for Digital Communications (IDCOM), Edinburgh University, Kings Buildings, Mayfield Road, Edinburgh EH9 3JL, UK
[b]CBR Devision, Dstl, SP4 0JQ, UK

ABSTRACT

Raman spectroscopy is a well-established spectroscopic method for the detection of condensed phase chemicals. It is based on scattered light from exposure of a target material to a narrowband laser beam. The information generated enables presumptive identification from measuring correlation with library spectra. Whilst this approach is successful in identification of chemical information of samples with one component, it is more difficult to apply to spectral mixtures. The capability of handling spectral mixtures is crucial for defence and security applications as hazardous materials may be present as mixtures due to the presence of degradation, interferents or precursors. A novel method for spectral unmixing is proposed here. Most modern decomposition techniques are based on the sparse decomposition of mixture and the application of extra constraints to preserve the sum of concentrations. These methods have often been proposed for passive spectroscopy, where spectral baseline correction is not required. Most successful methods are computationally expensive, *e.g.* convex optimisation and Bayesian approaches.

We present a novel low complexity sparsity based method to decompose the spectra using a reference library of spectra. It can be implemented on a hand-held spectrometer in near to real-time. The algorithm is based on iteratively subtracting the contribution of selected spectra and updating the contribution of each spectrum. The core algorithm is called fast non-negative orthogonal matching pursuit, which has been proposed by the authors in the context of nonnegative sparse representations. The iteration terminates when the maximum number of expected chemicals has been found or the residual spectrum has a negligible energy, i.e. in the order of the noise level. A backtracking step removes the least contributing spectrum from the list of detected chemicals and reports it as an alternative component. This feature is particularly useful in detection of chemicals with small contributions, which are normally not detected. The proposed algorithm is easily reconfigurable to include new library entries and optional preferential threat searches in the presence of predetermined threat indicators.

Under Ministry of Defence funding, we have demonstrated the algorithm for fingerprinting and rough quantification of the concentration of chemical mixtures using a set of reference spectral mixtures. In our experiments, the algorithm successfully managed to detect the chemicals with concentrations below 10 percent. The running time of the algorithm is in the order of one second, using a single core of a desktop computer.

Keywords: Raman Spectroscopy, Spectral Decomposition, Spectral Quantification and Fingerprinting

1. INTRODUCTION

Optical spectroscopy methods are based on recording the interaction of light with the materials. These interactions through absorption, transmission, reflection or scattering can provide fundamental, and in some instances characteristic, information from the material under study. Raman spectroscopy, in contrast to other spectroscopic techniques, is based on measuring the scattered light from a very narrow band illuminated light, *i.e.* a

Further author information: (Send correspondence to M.Y.)
M.Y.: E-mail: yaghoobi@ieee.org, Telephone: +44 (0)131 651 3492
D.W.: E-mail: d.wu@ed.ac.uk
R.J.C.: E-mail: rjclewes@mail.dstl.gov.uk
M.E.D.: E-mail: mike.davies@ed.ac.uk, Telephone: +44 (0)131 650 5795

laser. Raman spectroscopy has been shown to be an accurate technique for fingerprinting the unknown chemical molecules. A conventional method to subsequently identify the unknown chemicals is to compare the recorded spectrum with a reference library of chemicals. As this can be processing heavy, an alternative identification approach may be applied using the correlation of peak positions only. This method is however useful for the detection of few chemicals in a mixture.

A challenge associated with visible wavelength excitation Raman is that the recorded spectra may contain*; a) Raman spectral peaks and, b) noise fluorescence.[1] Due to the intensity of the fluorescence effect, underlying Raman spectral features may be sufficiently convolved with fluorescence to make automated correlation analysis approaches challenging. As a result this part of the incoming spectra is usually subtracted for automatic spectral fingerprinting. To practically demonstrate this fact, we have shown a sample set of a typical spectral library in Figure 1. On the left hand column, we have shown four spectra, which contain fluorescent spectral features. After removing the background, we get the middle column, with the background of each spectrum shown on the right hand plot. The difference between the feature rich form of the Raman spectra and broad nature of the background signals are clear.

The peak matching algorithms are extremely dependent on the baseline corrections, as the raw spectra may be dominated by the fluorescence. Various algorithms have been presented to compensate the fluoresence.[2–7] The baseline correction process establishes a logical relation between the contribution of each spectrum in the spectral mixture, and the actual concentrations of corresponding chemicals.

When the backgrounds of the spectral library and the mixture are removed, the mixture can be modelled as a combination of the library elements with an additive noise. While the combination of library elements can be generally distorted by a nonlinearity artefact, we practically observed that, a linear model can well describe a series of mixtures, while the most non-linear effects will occur with very polar and very polarisable mixture components. Such a model, which will be discussed in section 2, has been a basic foundation of spectral deconvolution methods for the other spectral modalities.[8,9] A key assumption in the approach presented here is that only a few library elements exist in sufficient quantity in unknown mixtures. In this setting, sparse representations are used to decompose the spectral mixture to elementary spectra, alongside noise†.[10] Promising results are derived by assuming such a model for spectral mixtures and using standard sparse approximation algorithms. Most sparse approximation algorithms are iterative (semi-) optimisation algorithms, which need to run for a particular number of iterations or achieve a satisfactory convergence criteria.[10] As a result, the computational cost of most algorithms are high. If the task is to reduce the computational cost, with the aim of running in a (close to) real-time application on a computationally/memory limited platform, we have to carefully choose the sparsity algorithm.

A class of low-complexity sparse approximation algorithms are greedy algorithms and the most commonly used is the Matching Pursuit (MP) algorithm.[11] The algorithm is based on gradually adding the most currently correlated library element to the set of selected elements, until the residual energy vanishes. There is a modified version of MP, which updates the coefficients related to the contribution of each library element at each iteration, called the Orthogonal Matching Pursuit (OMP) algorithm.[12] Recall that one expects to find positive contributions in a spectral decomposition problem. We can add the non-negativity constraint to the greedy selection process.[13] A Fast Non-Negative OMP (FNNOMP) has been presented by the authors in reference,[14] which can be efficiently used in the spectral decomposition, which will be presented later.

After sparse spectral decomposition, some post-processing steps are necessary to guarantee that the fingerprinting and quantification have been appropriately done. We will discuss about such process in Section 3.

Some real and synthetic mixtures have been used in the simulation Section 4 to demonstrate the capabilities of the proposed low-complexity algorithm for fingerprinting and quantification.

*If the sample is fluorescent at the illuminated wavelength, an alternate excitation wavelength may be used.

†Any model mismatch errors can be counted as noise term.

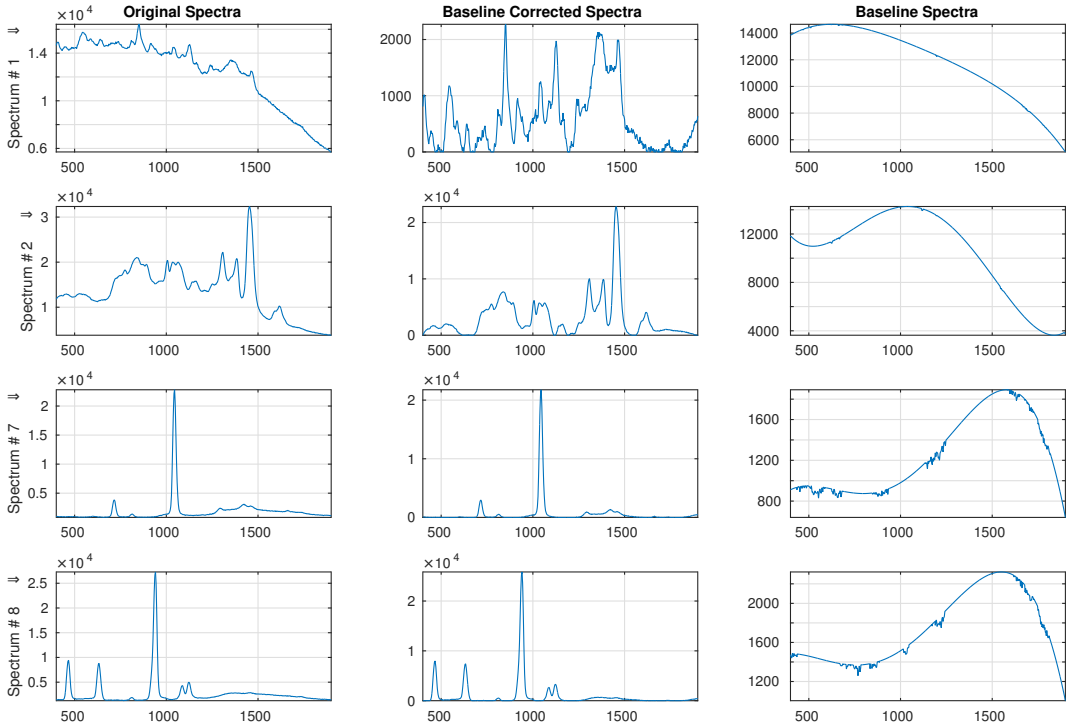

Figure 1. The spectral libraries; a) original spectra (left column), b) baseline corrected spectra (middle column) and c) baseline functions, which are found using a baseline correction algorithm (right column).

2. MATHEMATICAL FORMULATION

The advanced technologies in modern hand-held spectrometers, allow us to have low-noise accurate spectral measurements. We will model a typical spectral generation model in the next section, which has been used in the rest of paper. The model is simple and accurate for the purpose of fingerprinting and coarse quantification.

2.1 Signal Model

A set of reference spectra is assumed known, and characterised by library, $\mathcal{M}$. Such spectra can generate a matrix of spectral library $\mathbf{M}$, by putting the pure spectra $\mathbf{m}_j$'s in the columns of $\mathbf{M}$. We can characterise the contribution of each chemical by a "positive" coefficient α_j. Let the recorded spectral mixture be $\mathbf{y} \in \mathbb{R}^d$, where y_i is the measurement at ith wavenumber. The generation of $\mathbf{y}$ can be modelled as follows,

$$\mathbf{y} = g\left(\mathbf{M}, \boldsymbol{\alpha}\right) + \mathbf{b} + \boldsymbol{\omega}, \tag{1}$$

where function g can generally be a non-linear mixing model, $\boldsymbol{\alpha} = [\alpha_j]_{j=1:N}$, $\mathbf{b}$ is baseline signal and $\boldsymbol{\omega}$ includes the measurement noise and any unknown spectra. The unknown spectra are the signature of chemicals which are not included in the library. If the size of library is small the chance of an unknown spectra in $\boldsymbol{\omega}$ is usually higher. We do not investigate this scenario here, while the proposed approach can be modified to handle such a scenario.[10] The mixing model g includes a large class of nonlinear functions to describe the actual spectral mixing process, visible light hyperspectral unmixing analysis provides examples of this.[15,16] The nature of Raman spectral nonlinearities is different to those in other spectral modalities, and it is mainly caused by interaction between different chemicals in the mixture. Such a nonlinearity can change the locations of the spectral peaks. however, for the purposes of this study non-linear Raman mixtures will not be considered. The model will then be simplified as the following linear model,

$$\mathbf{y} = \mathbf{M}\boldsymbol{\alpha} + \mathbf{b} + \boldsymbol{\omega}. \tag{2}$$

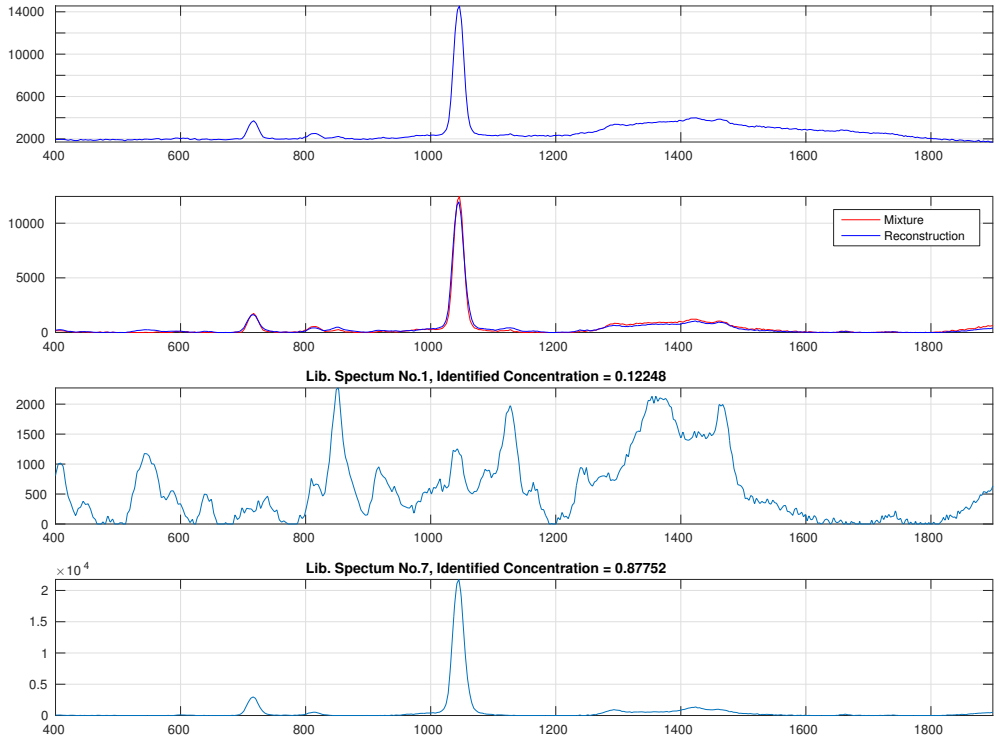

Figure 2. Spectral decomposition; mixture (top) baseline corrected mixture and reconstructed signals (second top), library elements # 1 and # 7 (bottom two). The ground truth for the concentrations of library components # 1 and # 7 are respectively 5% and 95%.

The linear approximation can be very useful as various signal processing techniques exist which operate on such a mixing model. We must, however, be careful with (2), specifically with regards to i) the relative magnitude of baseline signal $\mathbf{b}$ and ii) non-ideal mixing model $\mathbf{M}\boldsymbol{\alpha}$. The magnitude of baseline $\mathbf{b}$ can be larger than the other two terms, *i.e.* $\mathbf{M}\boldsymbol{\alpha}$ and $\mathbf{w}$. An approach to mixture decomposition, includes a baseline removal pre-processing to compensate the effect of large $\mathbf{b}$. In section 2.2, we briefly review some available baseline correction methods for Raman spectroscopy. A small non-linearity artifact can be handled as an extra additive error to the linear model, which can be absorbed into $\mathbf{w}$. This is only acceptable setting if the linear model is overally a good describing generating model. This is the case observed in our set of experiments[‡].

If the baseline of the measurement is removed, the spectral mixture model can be simplified as,

$$\mathbf{y} = \mathbf{M}\boldsymbol{\alpha} + \mathbf{w}, \tag{3}$$

where we assume that the model-mismatch and baseline removal algorithm residual are small and absorbed in $\mathbf{w}$. In a more involved case, the number of library elements may be more than the number of observation, *i.e.* $N > d$. The inverse problem in such a scenario is underdetermined. To resolve the issue and find the solution, we need extra priori information about $\boldsymbol{\alpha}$, which is sparsity here. The sparsity of the coefficient vector means that in a normal chemical mixture only a few component exist. However, directly solving (3) with the sparsity assumption for the $\boldsymbol{\alpha}$ may require checking all possible combinations of spectra to find the actual mixture. Even in the

[‡]This excludes the cases in which some chemical concentrations are very small. In such a setting, a more complicated model or an adaptive decomposition strategy is necessary

overdetermined setting $N < d$, as there are normally significant similarities between the spectra in the library, it is important to regularise the inverse problem. Such a similarity makes the decomposition more challenging, which will be discussed in section 2.3.

With a simple example, we now show that the linear model (3) is satisfactory for Raman spectral decomposition. We have shown a mixture spectrum of components number 1 and 7, with the concentrations 5% and 95%, respectively in Figure 2. This is a challenging unmixing scenario due to both the low Raman scattering efficacy and low concentration of component number 1. We show these spectra in the bottom two windows of Figure 2.

Many spectral decomposition techniques are based on solving a regularised inverse problem. The aim of regularisation is to make the overall decomposition robust to the model mismatch, imperfect baseline removal and measurement noise. The regularisation is usually enforced by penalising a metric of the coefficient vector $\boldsymbol{\alpha}$ or enforcing a low-dimensional structure like sparsity. The authors proposed a sparsity based Raman spectral decomposition in[10] and explained how it can help finding spectral decompositions in a small scale setting. Such an algorithm is computationally expensive, which prohibits its application in a computationally limited embedded system. We introduce a new technique of decomposing the spectra in section 3, to address this problem specifically.

In the described models, we assume that a library of pure spectra is given. The library for this study was collected using a hand-held Raman spectrometer to analyse pure chemicals ($> 95\%$ purity) and then baseline corrected. The library spectra collected were representative sample measurements, no attempts were made to assure a constant signal to noise ratio for all entries.

2.2 Baseline Correction

The baseline artefact is morphologically different to the Raman spectral features. The former is smooth and the latter is feature rich. This fact can be used to separate the background signal $\mathbf{b}$ from the rest of measurement. The most canonical approach is an "optimisation" based decoupling. In this setting, we need to minimise a metric $\psi(\cdot)$ [§], as follows,

$$b = \text{argmin}_{\tilde{b}\in\mathcal{B}} \sum_{j=1}^{d} \psi(y_i - \tilde{b}_i), \tag{4}$$

where $\mathcal{B}$ is the set of smooth functions. The most frequent choice for ψ is the quadratic function $\psi(x) = x^2$ and $\mathcal{B}$ to be low-order polynomials. This setting has a closed form solution to find b, which is the application of pseudoinverse of $\mathbf{P}$, the matrix of sampled low-order polynomial functions. One issue with such a simple baseline correction is that there is no guarantee to have positive spectra after baseline correction. Another is that all material spectra will be correlated to some extent to the low order polynomial bases. To also have the positivity constraint on the residual $y_i - \tilde{b}_i$, some iterative optimisation techniques have been proposed. A non-quadratic metric has also been used in the optimisation program (4), to non-uniformly penalise the large/small and positive/negative values.[4] For example, a typical metric can be as follows,

$$\psi(x) = \begin{cases} x^2 & if\ x < s \\ s^2 & otherwise \end{cases}$$

where s is the threshold value and it is usually small, *e.g.* $s = 0.01$.

Another approach for baseline correction is to regularise a "weighted" quadratic function with the roughness penalty function of the baseline as follows,[5,6]

$$b = \text{argmin}_{\tilde{b}} \sum_{j=1}^{d} w_i(y_i - \tilde{b}_i)^2 + \lambda_V \sum_{j=2}^{d} (\tilde{b}_i - \tilde{b}_{i-1})^2, \tag{5}$$

where w_i's are some non-negative weights and λ_V is the regularisation parameter. w_i's are selected to exclude the quadratic penalty in the peak locations, by having zero weights. However, the manual selection of the weights

[§]A "metric" is a function which characterises the distance between two points in a signal space. There may be some cases in which the "metric" $\psi(\cdot)$ is not a mathematically correct distance, but it has most properties of a distance.

is difficult and prone to error. An adaptive technique for changing the weights has been presented elsewhere by Zhang *et al.*[7]

Another class of baseline correction techniques is based on Bayesian model of the Raman spectra and the background.[2,3] Here the feature rich part of the spectrum is modelled with a sum of Gaussian function and the background with a smooth function like cubic B-spline. As a direct evaluation of the mixture distribution is not possible, a stochastic numerical Bayesian method a Markov Chain Monte-Carlo method, has been proposed to estimate the joint distribution.[3] Such methods are computationally expensive and therefore not suitable for our purpose.

2.3 Coherence of the Library

A crucial factor in the success of the sparse approximation methods is called the coherence μ. It measures the maximal similarity between two distinctive elements. If $\mathbf{m}_j$ represents the jth column of the library matrix $\mathbf{M}$, the coherence is the maximum correlation between normalised $\mathbf{m}_j$ and $\mathbf{m}_k$, where $j \neq k$. μ can also be formulated as follows:

$$\mu = \max_{j,\ k} \frac{|\mathbf{m}_j^T \mathbf{m}_k|}{\|\mathbf{m}_j\|_2 . \|\mathbf{m}_k\|_2},$$

where ^{T} indicates the transpose of the vector and $\|\cdot\|_2$ indicates the Euclidean norm. A library with a large μ has some elements, which are very similar. In the sparse approximation terminology, the discovery of the correct sparsity support, *i.e.* the footprint of the chemicals, becomes more challenging. As the spectral library only lives in one out of 2^d possible orthants, the correlated between spectra can be fundamentally larger. This means that μ is often higher than generally positioned or other signed libraries. To demonstrate this fact, we used the nine baseline corrected Raman spectra, which some of them are shown in Figure 1. We calculated the coherence matrix, which has the correlation between $\frac{\mathbf{m}_j}{\|\mathbf{m}_j\|_2}$ and $\frac{\mathbf{m}_k}{\|\mathbf{m}_k\|_2}$ on the (j,k)-th location. The coherence matrix is shown in Figure 3. This figure shows that some elements are highly correlated, such that the correlation for some off-diagonal elements is larger than 0.9. This is an evidence of the difficulty of Raman spectral decomposition.

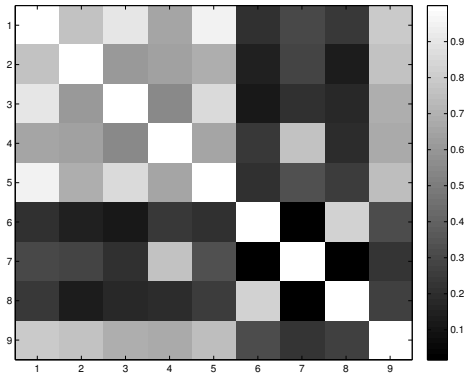

Figure 3. The coherence matrix of an example of a small Raman spectral library.

3. PROPOSED ALGORITHM

The task is to reliably solve the inverse problem of (3). For reasons explained in the introduction, such a set of linear equations is underdetermined and/or noise sensitive and requires some form of regularisation. The most frequent regularisation is the Euclidean norm $\|\boldsymbol{\alpha}\|$ to bound the energy of the decomposition coefficients. More precisely, we have to solve the following optimisation program,

$$\boldsymbol{\alpha} = \text{argmin}_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{M}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2,$$

where λ is a positive (Lagrange) multiplier. Solving this optimisation program is straightforward and can be done by application of pre-computed inverse of $\mathbf{M}^T\mathbf{M} + \lambda\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, *i.e.* $\boldsymbol{\alpha} = (\mathbf{M}^T\mathbf{M} +$

$\lambda\mathbf{I})^{-1}\mathbf{M}^T\mathbf{y}$. However, as there are no sparsity and non-negativity constraints on $\boldsymbol{\alpha}$, the solution normally has contributions from many library elements with possibly negative values. This is not representative of the spectral contributions. If we apply the non-negativity constraint, the problem does not have a closed form solution, and it is normally solved using iterative algorithms. An alternative regularisation is a sparsity inducing penalty function to consider the fact that we only expect to see a few chemicals, *e.g.* $k \ll N$ out of N chemicals, present in any one experiment, which will be discussed in the next section.

3.1 Sparsity Based Fingerprinting and Quantification

Some regularisation functions induce sparsity. The reason is that the minima of regularised cost functions with such penalty functions, are sparse or only a few elements have significant magnitudes, *i.e.* "compressible". The most popular class of sparsity functions is the p-"norm"'s $\|\boldsymbol{\theta}\|_p := (\sum_{j=1}^{N} |\theta_j|^p)^{\frac{1}{p}}$, $0 < p \leq 1$, in which we use $\|\cdot\|_p^p$ as the sparsity promoting regularisation. Among them, only $\|\cdot\|_1$ is convex. If we know the magnitude level of the error vector $\mathbf{w}$, $\|\mathbf{w}\|_2 \leq \epsilon$, a convex sparsity approach is to solve the following regularised optimisation problem,

$$\boldsymbol{\alpha} = \text{argmin}_{\boldsymbol{\theta} \dot{\geq} 0} \sum_j |\theta_j| \;\text{ s. t. }\; \|\mathbf{y} - \mathbf{M}\boldsymbol{\theta}\|_2 \leq \epsilon,$$

where $\dot{\geq}$ is an elementwise comparison operator. This optimisation can be reformulated as follows,

$$\boldsymbol{\alpha} = \text{argmin}_{\boldsymbol{\theta} \dot{\geq} 0} \frac{1}{\lambda_\epsilon} \|\mathbf{y} - \mathbf{M}\boldsymbol{\theta}\|_2^2 + \sum_j |\theta_j|, \tag{6}$$

where λ_ϵ is a regularisation multiplier and a function of noise level ϵ, which controls the level of sparsity in the final representation $\boldsymbol{\alpha}$. The formulation (6) is often preferred for the available wide range of optimisation techniques. The optimisation program (6) generates a sparse or compressible coefficient vector $\boldsymbol{\alpha}$, which is neither guaranteed to be exactly k-sparse nor has k dominant coefficients.

The next step of the algorithm is to identify the fingerprint of the chemicals, based on the coefficient vector $\boldsymbol{\alpha}$. This step is to project it onto the set of k-sparse vectors. This can be done by keeping the largest k components and shrinking the rest to zero. This is the case when we know the number of mixture components. Otherwise, we have to decide on the number of components in the mixture, based on the magnitude of the sorted elements of $\boldsymbol{\alpha}$.[10] We can then have a set $\mathcal{J}$ of non-zero elements.

For the relative quantification of components, as it is also discussed in,[17] we normalise $\boldsymbol{\alpha}$ to $\bar{\boldsymbol{\alpha}}$ such that $\sum_{j\in\mathcal{J}} \bar{\alpha}_j = 1$. The concentration percentage of each component j indexed in $\mathcal{J}$ can be simply calculated as $100 * \bar{\alpha}_j$.

The main challenge with sparsity based algorithms for embedded platforms is the computational load and the memory usage. Sparsity based optimisation programs, including (6), do not have a closed form solution and all available optimisation algorithms are iterative. The computational cost thus scale quadratically with the size of problem. Another issue with iterative convex optimisation algorithms is that it is not clear how many iterations are necessary for the algorithm to converge, which can compromise real-time implementations. In the following section, we propose a new low-complexity greedy sparse Raman spectral decomposition algorithm.

3.2 Low Complexity Greedy Sparse Spectral Decomposition

A non-negative sparse solution for (3) can be found using a non-negative MP-type method. Let α_k be the coefficient vector at stage k, which is initialised as a zero vector. The idea is to efficiently reduce the residual error $\mathbf{r}_k := \mathbf{y} - \mathbf{M}\boldsymbol{\alpha}_k$, by adding a new spectrum to the set of indexed spectra by $s = \text{supp}(\boldsymbol{\alpha}_k)$, where $supp(\cdot)$ is the support (the indices of non-zero elements) operator. The algorithm then finds the library spectrum that is most positively correlated to to the residual $\mathbf{r}_k$ at the kth iteration. To update the coefficient vector, NNOMP solves a least-square problem with a non-negative constraint, using the selected spectra at each iteration. A pseudocode of the Non-Negative OMP (NNOMP) is presented in Algorithm 1. In this algorithm, k indicates the iteration number, $\max(\cdot)$ in line 3 is the maximum operator with ζ as the maximum value and ι as the index of element with the maximum value, and s or $|_s$ indicates the selected sub-vector or sub-matrix, *i.e.* here selected

1: **initialisation:** $s = \emptyset$, $k = 0$, $\boldsymbol{\alpha}_0 = \mathbf{0}$ and $\mathbf{r}_0 = \mathbf{y}$
2: **while** $k < K$ & $\max(\mathbf{M}^T\mathbf{r}_k) > 0$ **do**
3: $(\zeta, \iota) \leftarrow \max(\mathbf{M}^T\mathbf{r}_k)$
4: $s \leftarrow s \cup \iota$
5: $\boldsymbol{\alpha}_s \leftarrow \text{argmin}_{\boldsymbol{\theta} \geq 0} \|\mathbf{y} - \mathbf{M}_s\boldsymbol{\theta}\|_2$
6: $\mathbf{r}_{k+1} \leftarrow \mathbf{y} - \mathbf{M}_s\boldsymbol{\alpha}_s$
7: $k \leftarrow k + 1$
8: **end while**
9: $\boldsymbol{\alpha}|_s \leftarrow \boldsymbol{\alpha}_s$

Algorithm 1: Canonical Non-Negative Orthogonal Matching Pursuit

columns, by the index set s. The termination condition for such an iterative algorithm can be the sparsity of the representation, which is the number of mixture components, or the maximum energy which can be absorbed using a new library element, related to the system noise. In this case, if $\max(\mathbf{M}^T\mathbf{r}_k)$ is small, the new component is unlikely to contribute to the mixture and we thus stop the algorithm.

Such an implementation for NNOMP algorithm can be counted as the "canonical" version, *i.e.* slow implementation. The computational complexity is mainly in the calculation of Non-Negative Least Square (NNLS) step 5, which does not have a closed form or computationally light-weight solution. As a result, the authors developed an alternative fast version.[14] The fast NNOMP is based on the QR factorisation of the library sub-selected matrix $\mathbf{M}_s$, *i.e.* a matrix generated only using the indexed columns by s. To guarantee the non-negativity of the coefficient vector, we sometimes need to ignore the most positively correlated library spectrum and select another good candidate. The details of the FNNOMP has been presented in.[14]

The fast sparsity Raman spectral decomposition algorithm uses the FNNOMP implementation with the termination condition for the sparsity method as the residual energy $\mathbf{r}_k$, with a backtracking step, if the last selected spectrum has a small contribution. In this case we indicate this spectrum as an alternative component, which can be useful for identifying small traces of the chemicals. The algorithm is fast and is free of any iterative sub-task, *e.g.* NNLS optimisation. Loading and storing the Raman spectral library is the most memory intensive part of the proposed algorithm.

Note that the proposed algorithm is working with the formulation of (3), which needs a baseline correction as the preprocessing step. As the baseline correction of the library spectra can be done off-line, the main computational burden of the whole process is the baseline correction of the input spectrum.

4. SIMULATION RESULTS

The simulation results will be presented in two sections, based on synthetic and real mixtures. A sub-set of a nine-element library is presented in Figure 1. The library element 9 is related to the instrument internal noise spectra initially included in the library so as to challenge the algorithm with a signature likely to be common to all spectra, this was subsequently removed from the mixture generation process. The synthetic simulations enable us to investigate the performance of algorithm with respect to the effect of different factors, *e.g.* the minimum concentration level of a chemical and different background artefacts.

γ	0.05	0.10	0.15	0.20	0.25
Ex. Rec. %	85.77	87.53	88.16	88.68	89.45

Table 1. The percentages of correct fingerprinting, with respect to the possible minimum concentration values γ.

4.1 Synthetic Mixtures

We can generate a synthetic mixture by adding library components. The linear mixture generation can be done using background compensated spectra and adding a typical background signal. We excluded the 9th library element and linearly combined $K = 2$ spectra with non-negative weights, which are chosen to be larger than a threshold θ and the sum to be equal to 1. We randomly selected 2 out of $N = 8$ library elements, combined

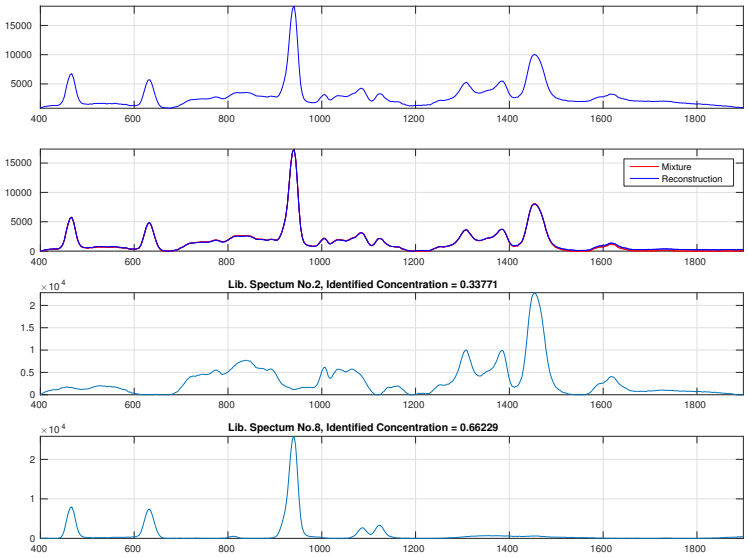

Figure 4. Spectral decomposition: synthetic mixture (top), baseline corrected mixture and reconstructed signals (second top) library candidates # 1 and # 2 (bottom two). The ground truth concentrations for the components # 1 and # 2 are respectively 34.37% and 65.63%.

the selected spectra with randomly generated concentrations and a randomly selected baseline based on the example library data set as discussed in Figures 1 and 2. Such baseline signals were generated by subtracting the baseline corrected signals from the original library elements. A typical mixture is shown in the top window of Figure 4. The second window shows two curves, related to the baseline corrected spectra and the reconstructed spectra using the spectral decomposition method. The two bottom windows show the selected spectra for this decomposition, which are correctly identified.

By randomly generating a large number of such synthetic mixtures, we measure the rate of successful fingerprinting. We here demonstrate the success rate of the algorithm for the mixtures of $K = 2$ components whilst changing the minimum possible concentrations. The success rate, averaged over 10000 trials, is shown in Table 1, where $\gamma < 1$ is the minimum possible concentration. As γ becomes larger, the unmixing is more successful. The reason is that the energy of the spectral mixture is more equally distributed between different components, which facilitates the correct detection of chemicals.

4.2 Physical Mixtures

We now present results on physical mixtures, with some analysis was performed blind, in the absence of ground truth information. The dataset has multiple mixtures, which we only know the ground truth about two mixtures, labelled “i” and “iv”. Mixture “i” and its decomposition output is shown in Figure 2. The output figures of other mixtures are presented in Table 2. The alternative components are presented in parenthesis. We have also shown the decomposition outputs of a standard sparsity based decomposition method,[10] for the comparison. The fingerprints and concentrations are similar, while there exist some mixtures with slightly different results. Mixtures “i” and “iv” are both the mixtures of chemicals #1 and #7, with the respective concentrations of 5%/95% and 20%/80%.

The aim here is to reduce the complexity of decomposition algorithm. We measured the running time of Matlab implementations of proposed and reference algorithms, using a single core of a desktop. The averaged running time of the decomposition of 11 different mixtures were respectively 11.05 and 1.09 seconds, which shows an acceleration factor of 11.

Mixtures	Predicted Fingerprints		Predicted Concentrations (in %)	
	Standard	Fast	Standard	Fast
i	1,7,(9)	1,7,(9)	[12 88]	[12 88]
ii	1,7,(9)	1,7,(9)	[17 83]	[17 83]
iii	1,7,(3)	1,7,(9)	[20 80]	[20 80]
iv	1,7,(9)	1,7,(9)	[44 56]	[44 56]

Table 2. Spectral recovery of physical mixtures. Here, the indices in parentheses represent alternative components. 9 is instrument noise and 3 is an analogue of 1.

5. CONCLUSION

Raman spectral decomposition can be used for chemical fingerprinting and relative quantification. The generative model of the spectra can be complicated, but sparse approximation techniques help to decompose the spectra, with a high level of accuracy. We presented a fast non-negative sparse approximation method for computationally limited hand-held Raman spectrometers. The proposed algorithm is based on an iterative selection of the new spectral components, from a reference library. The decomposition, after some post-processing, can be used for fingerprinting and rough quantification.

Our experiments demonstrated that the baseline correction of the input spectra is the dominant computational component. Some investigation on the efficient methods for the real-time applications of Raman spectral decompositions, is left for the future work.

The fast Raman spectral decomposition algorithm has been tested with other library sizes, which we only demonstrated the application to small libraries here. The fingerprinting with large libraries can be challenging, as more reference spectra means higher correlations between the spectra. The challenge in the application of small libraries, is the need to detect and remove any "unknown" components, which has been briefly investigated.[10] The full investigation of the effect of library size, has been left for the future work.

ACKNOWLEDGMENTS

This work is was supported in part by the University Defence Research Collaboration (UDRC) for signal processing in the networked battlespace, number EP/K014277/1, EPSRC platform grant, number EP/J015180/1 and Dstl enabling contracts ED-TIN2-4 and ED-TIN2-6.

REFERENCES

[1] Ferraro, J. R., [*Introductory raman spectroscopy*], Academic press (2003).

[2] Fischer, R., Hanson, K., Dose, V., and von Der Linden, W., "Background estimation in experimental spectra," *Physical Review E* **61**(2), 1152 (2000).

[3] Razul, S. G., Fitzgerald, W., and Andrieu, C., "Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **497**(2), 492–510 (2003).

[4] Mazet, V., Carteret, C., Brie, D., Idier, J., and Humbert, B., "Background removal from spectra by designing and minimising a non-quadratic cost function," *Chemometrics and Intelligent Laboratory Systems* **76**(2), 121 – 133 (2005).

[5] Cobas, J. C., Bernstein, M. A., Martn-Pastor, M., and Tahoces, P. G., "A new general-purpose fully automatic baseline-correction procedure for 1d and 2d NMR data," *Journal of Magnetic Resonance* **183**(1), 145–151 (2006).

[6] Zhang, Z.-M., Chen, S., Liang, Y.-Z., Liu, Z.-X., Zhang, Q.-M., Ding, L.-X., Ye, F., and Zhou, H., "An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy," *Journal of Raman Spectroscopy* **41**(6), 659–669 (2010).

[7] Zhang, Z.-M., Chen, S., and Liang, Y.-Z., "Baseline correction using adaptive iteratively reweighted penalized least squares," *Analyst* **135**(5), 1138–1146 (2010).

[8] Iordache, M., Bioucas-Dias, J., and Plaza, A., "Sparse unmixing of hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on* **49**(6), 2014–2039 (2011).

[9] Qian, Y., Jia, S., Zhou, J., and Robles-Kelly, A., "Hyperspectral unmixing via sparsity-constrained nonnegative matrix factorization," *Geoscience and Remote Sensing, IEEE Transactions on* **49**(11), 4282–4297 (2011).

[10] Wu, D., Yaghoobi, M., Kelly, S., Davies, M., and Clewes, R., "A Sparse Regularized Model for Raman Spectral Analysis," in [*Sensor Signal Processing for Defence*], (2014).

[11] Davis, G., Mallat, S., and Zhang, Z., "Adaptive time-frequency decompositions with matching pursuit," in [*SPIE's International Symposium on Optical Engineering and Photonics in Aerospace Sensing*], 402–413, International Society for Optics and Photonics (1994).

[12] Pati, Y., Rezaiifar, R., and Krishnaprasad, P., "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in [*Asilomar Conference on Signals, Systems and Computers*], 40–44 (1993).

[13] Bruckstein, A., Elad, M., and Zibulevsky, M., "Sparse non-negative solution of a linear system of equations is unique," in [*3rd International Symposium on Communications, Control and Signal Processing, ISCCSP*], 762–767 (2008).

[14] Yaghoobi, M., Wu, D., and Davies, M., "Fast Non-Negative Orthogonal Matching Pursuit," *IEEE Signal Processing Letters* **22**(9) (2015).

[15] Halimi, A., Altmann, Y., Dobigeon, N., and Tourneret, J.-Y., "Nonlinear unmixing of hyperspectral images using a generalized bilinear model," *IEEE Transactions on Geoscience and Remote Sensing* **49**(11), 4153–4162 (2011).

[16] Altmann, Y., Dobigeon, N., McLaughlin, S., and Tourneret, J.-Y., "Nonlinear spectral unmixing of hyperspectral images using Gaussian processes," *IEEE Transactions on Signal Processing* **61**(10), 2442–2453 (2013).

[17] Afseth, N. K., Segtnan, V. H., and Wold, J. P., "Raman Spectra of Biological Samples: A Study of Preprocessing Methods," *Applied Spectroscopy* **60**, 1358–1367 (Dec. 2006).